IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:

Sarah H. BASSON, et al.

Serial 10/058,143
No.:
Confirmation No.:7380

Filed. January 29, 2002

Docket No.:P27050

Group Art Unit:2655

Examine: Brian L. Albertalli

For: **COLLABORATION OF MULTIPLE AUTOMATIC SPEECH RECOGNITION (ASR) SYSTEMS**

United States Patent and Trademark Office
Randolph Building
401 Dulany Street
Alexandria, VA 22314

DECLARATION UNDER 37 C.F.R. § 1.131

Sir:

We, Sarah H. Basson, Dimitri Kanevsky and Emmanuel Yashchin do hereby declare:

1. We are co-inventors of the subject matter disclosed and recited in independent claims 1, 14 and 19 of the above-identified application.

2 We completed the invention of claims 1, 14 and 19 (and those claims dependent thereon) in the United States before September 7, 2001, as evidenced below.

CONCEPTION

3. Before September 7, 2001, we conceived of a method of integrating acoustic data using speech recognition, a system for integrating acoustic data using speech recognition and a machine readable medium containing code for integrating acoustic data

using speech recognition as disclosed and recited in claims 1, 14 and 19 of the application, an embodiment of which is evidenced by IBM Invention Disclosure YOR8-201-0381 (hereinafter referred to as "the Invention Disclosure") attached hereto as Exhibit A, in addition to other related documents. The Invention Disclosure and other related documents attached hereto is a photocopy of and is identical to the originals, except that all pertinent dates have been removed therefrom.

4. All dates removed from the Invention Disclosure and other attached documents attached hereto are before September 7, 2001.

5. As evidenced in the attached documents including the Invention Disclosure, the inventors conceived of and reduced to practice the following inventive features:

A. A method of integrating acoustic data using speech recognition including detecting voice data on a first computer and at least a second computer. The method includes identifying the voice data as a first master speaker associated with a speech recognition system residing on the first computer and providing the voice data of the first master speaker, from the first computer, to at least the second computer having a speech recognition system residing thereon. The method also includes analyzing the voice data residing on the first computer and at least the second computer, and integrating the analyzed voice data from the first computer and at least the second computer into a single decoding output.

B. A system for integrating acoustic data using speech recognition. A communication module which receives voice data from a plurality of computers each has speech recognition residing thereon, along with the communication module residing on the plurality of computers or a remote server. The system includes an evaluator module associated with each of the plurality of computers, as well as the evaluator module analyzing the voice data from each of the plurality of computers. Further still, the system includes an integrator module associated with the evaluator module, and integrates all of the analyzed voice data from each of the plurality of computers and provides one decoding output.

C. A machine readable code includes detecting a second voice data on the first computer and at least the second computer, and identifies the second voice data as being a second master speaker associated with the speech recognition system of at least the second computer. The machine readable code provides the second voice data to the first computer from the at least second computer, and analyzes the second voice data residing on the first computer and at least the second computer. Further still, the machine readable code integrates the analyzed first voice data and the analyzed second voice data into the single decoding output.

7. The benefits and features of a method of integrating acoustic data using speech recognition, a system for integrating acoustic data using speech recognition and a machine readable medium containing code for integrating acoustic data using speech recognition, are shown and described in the Invention Disclosure and accompanying documents.

8. These features and others are exemplified in the accompanying figures of the Invention Disclosure.

DUE DILIGENCE

9. Inventor Basson communicated with patent counsel in preparing a patent application based on the Invention Disclosure.

10. The inventors communicated with patent counsel in preparing a patent application for the above matter prior to September 7, 2001. We worked diligently on the preparation of the patent application until it was finalized for filing on January 29, 2002.

11. As an example, Mr. Calderon, of McGuireWoods, contacted the inventors, and more specifically, Mr. Yashchin prior to September 7, 2001 to discuss this invention and the preparation of a patent application. A draft application was forwarded to Mr. Yashchin on June 11, 2001, via facsimile. Further revisions were made and a final application was forwarded to the inventors, after working diligently on the draft applications, on December 28, 2001.

12. The final application was approved by the inventors and the inventors signed a declaration on January 2, 2002. The final draft of the patent application was filed in the U.S. Patent and Trademark Office on January 29, 2002.

13. We declare that all statements made herein of our own knowledge are true and that all statements made on information and belief are believed to be true; and further, that the statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.

Sara H Basson
Sarah H. Basson

4/14/05
Date

Dimitri Kanevsky

4/14/05
Date

Emmanuel Yashchin

Apr 14, 05
Date

(P27050.A03)

YOR92001 0346

Disclosure YOR8-2001-0381

Prepared for and/or by an IBM Attorney - IBM Confidential

Created By: Dimitri Kanevsky **Created On:**
Last Modified By: Dimitri Kanevsky **Last Modified On:**

Required fields are marked with the asterisk (*) and must be filled in to complete the form .

***Title of disclosure (in English)**

Colloboration of multiple ASR systems

Summary

Status	Under Evaluation
Processing Location	YOR
Functional Area	900 Goyal-Systems & Software
Attorney/Patent Professional	Stephen C Kaufman/Watson/IBM
IDT Team	Stephen C Kaufman/Watson/IBM
Submitted Date	
Owning Division	RES
Incentive Program	
Lab	
Technology Code	
PVT Score	40

Inventors with Lotus Notes IDs

Inventors: Sara H Basson/Watson/IBM, Dimitri Kanevsky/Watson/IBM, Emmanuel Yashchin/Watson/IBM

Inventor Name	Inventor Serial	Div/Dept	Inventor Phone	Manager Name
Basson, Sara H.	111773	22/3C1A	862-1270	Snayd, Paul F.
Kanevsky, Dimitri	202817	22/S1GA	862-2834	Gopinath, Ramesh A.
> Yashchin, Emmanuel	409501	22/X8MA	862-1828	Jensen, David L.

> denotes primary contact

Inventors without Lotus Notes IDs

IDT Selection

Select Functional Area

IDT Team: Stephen C Kaufman/Watson/IBM	**Attorney/Patent Professional:** Stephen C Kaufman/Watson/IBM

Response Due to IP&L :

***Main Idea**

1. Describe your invention, stating the problem solved (if appropriate), and indicating the advantages of using the invention.

it is necessary to create protocols of manymeetings. Manual protocolling is expensive and not always available. Individual ASRs do not have sufficient quality to provide the protocols.

2. How does the invention solve the problem or achieve an advantage,(a description of "the invention", including figures inline as appropriate)?

1. Every speaker has a processor associated with him (can be in his laptop) that is capable of

(a) Identifying his "master", i.e. being able to filter out signal corresponding to person he is associated with from the environment
(b) Being able to recognize what his "master" said (possibly with the assistance of topic identification, environment identification, tracking number of speakers present or other techniques)
(c) Presenting to the Referee the statement of type: (My master said: "It came with my pea sea") and I associate two numbers (both between 0 and 1) with this staement:

0.99 score that it was my "master" who said it
0.60 score that the statement was recognized correctly

(d) Receive from the Referee feedback about its performance
(e) When not recognizing their "master", maintain its own record of speakers and text, and being able to present it to Referee (automaticaly, or upon request by the Referee).

The act of a user processor presenting his version of text to the Referee is called a "bid". Note that not all processors need to run the same speech recognition program: for example, some could run Dragon, others - VV.

The Referee program is responsible for maintaining a stenographic record of the conversation (including cases where 2 speakers are talking simultaneously); it must be able to

(i) Receive "bids" from individual processors
(ii) Decide which "bids" will be accepted into official text record (this record is available to participating processors), and what text needs to be corrected; for example, it could accept the above claim about the identity of the speaker, but enter corrected version: "It came with my PC" into official record
(iii) Notify individual processors on disposition of their "bids" and introduced corrections.
(iv) Maintain a record of "credibility" of various processors on their ability to recognize their master and the text. This record can be used to adaptively improve Referees performance (for example, Referee could find one of the programs so unreliable that it gives this processor a credibility index of 0 and puts in its own version of speaker/text, possibly after polling other processors for their version of the speaker/text). In other words, the "good" processors could help the Referee to maintain the record, even when some of the processors are bad. The credibility record can also be used by individual processors to improve their performance

3. If the same advantage or problem has been identified by others (inside/outside IBM), how have those others solved it and does your solution differ and why is it better?

4. If the invention is implemented in a product or prototype, include technical details, purpose, disclosure details to others and the date of that implementation.

***Critical Questions (Questions 1-9 must be answered)**

(no hyphens)

Stephen C Kaufman

YOR9- .0346

To: Kathy Cognatello/Watson/IBM@IBMUS
cc:
From: Stephen C Kaufman/Watson/IBM@IBMUS
Subject: *IBM Confidential: IBM Docket No.

---------- Forwarded by Stephen C Kaufman/Watson/IBM on 04/13/2001 08:56 AM ----------

Dimitri Kanevsky

To: Stephen C Kaufman/Watson/IBM@IBMUS
cc: Sara H Basson/Watson/IBM@IBMUS, Emmanuel Yashchin/Watson/IBM@IBMUS
From: Dimitri Kanevsky/Watson/IBM@IBMUS
Subject: *IBM Confidential: IBM Docket No.

Stephen,

Please forward to C.Lamont Whitham the fuller patent description that is attached here.
I will send you figures. Thanks,
Dimitri

multipleASR.lw

[redacted]

. IBM Docket

Colloboration of multiple ASR systems

Collective Usage of a Speech Recognition Machine For Improved Transcription

Sara Basson, Dimitri Kanevsky, Emmanuel Yashchin

Prior Art

The transcription of meetings is a very important application. At present, the transcription of meetings is done through either, stenography or some other person that records the main points of a meeting, or when several people meet to discuss a particular topic. These methods are not ideal because a stenographer may not be available, or may be too expensive; similarly, a summary of a meeting or discussion may skip over many important details, and may also be expensive. The use of speech recognition to transcribe a meeting and compose a summary is rather difficult because of the high error rate. A summary based on text collected by speech recognition is also difficult.

Summary

Our invention is based on the idea that a majority of people attending meetings bring their own laptops. These people should have a speech recognition system installed in their laptops. The speech recognition machine/system should be trained for the user. Each speech recognition machine must run an application that allows all of the speech recognition systems to cooperate amongst themselves. There may be a general computer or laptop that is used to coordinate the rest of the laptops. When each user speaks at the meeting the speech recognition systems must cooperate with each other by, first of all, recognizing their own master, and then sending the decoding to the central server/referee, which is also receiving and evaluating information that it is receiving from other speech recognition machines. Finally, the speech recognition server chooses the best resulting transcription on the basis of the information that it receives from the many laptops at the meeting.

The speech recognition systems also send voice data or results of signal processing data from other speech recognition machines to the central server/referee. Therefore, the machines located at a distance from the speaker may also participate in the decoding process. Parallel decoding on several machines improves the algorithms produced from parallel speech recognition systems. One of the methods that allows for improving speech recognition is called 'Rover', a voting system that chooses the most frequent set of similar decoded text from many entries by several speech recognition systems. For example if 5 speech recognition machines chose one word, and three speech recognitions systems chose another word, then the system assumes that the word chosen by the 5 machines was the correct word.

1. Describe your invention, stating the problem solved (if appropriate), and indicating the advantages of using the invention.

[redacted]

it is necessary to create protocols of manymeetings. Manual protocolling is expensive and not always available. Individual ASRs do not have sufficient quality to provide the protocols.

2. How does the invention solve the problem or achieve an advantage,(a description of "the invention", including figures inline as appropriate)?
1. Every speaker has a processor associated with him (can be in his laptop) that is capable of

(a) Identifying his "master", i.e. being able to filter out signal corresponding to person he is associated with from the environment
(b) Being able to recognize what his "master" said (possibly with the assistance of topic identification, environment identification, tracking number of speakers present or other techniques)
(c) Presenting to the Referee the statement of type: (My master said: "It came with my pea sea") and I associate two numbers (both between 0 and 1) with this staement:

0.99 score that it was my "master" who said it
0.60 score that the statement was recognized correctly

(d) Receive from the Referee feedback about its performance
(e) When not recognizing their "master", maintain its own record of speakers and text, and being able to present it to Referee (automaticaly, or upon request by the Referee).

The act of a user processor presenting his version of text to the Referee is called a "bid". Note that not all processors need to run the same speech recognition program: for example, some could run Dragon, others - VV.

The Referee program is responsible for maintaining a stenographic record of the conversation (including cases where 2 speakers are talking simultaneously); it must be able to

(i) Receive "bids" from individual processors
(ii) Decide which "bids" will be accepted into official text record (this record is available to participating processors), and what text needs to be corrected; for example, it could accept the above claim about the identity of the speaker, but enter corrected version: "It came with my PC" into official record
(iii) Notify individual processors on disposition of their "bids" and introduced corrections.
(iv) Maintain a record of "credibility" of various processors on their ability to recognize their master and the text. This record can be used to adaptively improve Referees performance (for example, Referee could find one of the programs so unreliable that it gives this processor a credibility index of 0 and puts in its own version of speaker/text, possibly after polling other processors for their version of the speaker/text). In other words, the "good" processors could help the Referee to maintain the record, even when some of the processors are bad. The credibility record can also be used by individual processors to improve their performance

3. If the same advantage or problem has been identified by others (inside/outside IBM), how have those others solved it and does your solution differ and why is it better?

4. If the invention is implemented in a product or prototype, include technical details, purpose, disclosure details to others and the date of that implementation.

Brief description of figures

Detailed Description of Figures

Figure 1 represents 3 users, user 100, user 101 and user 102. Modules 104,105, and 106 represent the laptops held by each user. Modules 107,108, and 109 represent microphones. Module 110 is the server/cpu that runs the cpu program. The laptops 104,105,and 106 are connected to each other and the cpu via wireless network. When user 100 speaks, and users 101 and 102 are silent, the cpu 104 decides that it their master that is speaking and not user 101 or 102. Simultaneously cpu's 105 and 106 determine that their masters are not speaking. But, the microphones pick up the voice of the speaker. The closer microphones pick up the speaker with better clarity and increased volume. Using this factor the cpu's are able to determine the approximate distance of the speaker and therefore determine if the speaker is their master. If the cpu determines that their master is speaking then the voice in microphone is send through another driver from laptop 104 to 105 to 106 . Driver 141 receives sound input from microphone 107 and transmits the data to the speech recognition system in laptop 105. Similarly, driver 142 receives sound data from microphone 107 and transmits this information to the ASR system in computer 106. The Automatic Speech Recognition systems are represented by modules 130, 131, and 132. Voice data may be sent to and from each laptop through a communication module or through the server 110. Each speech recognition recognizes the voice of their speaker and sends this information to the referee program 120 to produce a better decoding. The method for producing a better decoding is described in a later figure. All the laptops are constantly monitoring that their master has not begun talking. As soon as it is determined that another person has begun speaking, the situation has changed and, for example the data is sent from user 101 to users 102 and 100.

Figure 2 contains communication module 200 that receives and sends information to and from clients. Decoding data 201 from the laptops is sent through the communication module. The decoder data may be composed of decoding data from each laptop, i.e. decoding 1-202, decoding 2- 203, and decoding n(from laptop n)- 204. All of this data is sent through the evaluator of each decoder output 207. The decoder data is analyzed and receives a confidence score. If a confidence measure is calculated for each decoder output, what is the chance that the word was placed correctly- this results in a likelihood score. Therefore, the decoder data consists of the word and the likelihood score of the the word. The confidence score may be assigned in the local laptops of the users and may be sent to the referee program. The evaluator of each output can then rely on receiving a higher level language model. Language model 205 may be used to determine the chance of each type of text, evaluate the perplexity of a given text, and determine chance of the proper word being placed correctly amidst the remeinder of the text. It is assumed that the ASR programs are different, because people may have different ASR programs in their respective laptops. Even in those cases where people have identical ASR programs they may have different decoding methods. A detailed description of using different decoding methods with one ASR will be discussed in a later figure. In this way, each speech recognition produces a text that is variable from the text of other ASR systems. Module 208 integrates all of the decoder data from all of the ASR systems into one decoding output. For example, a Rover method is utilized according to reference number XXXXXX. This is based on a voting system that chooses the word that was chosen by the majority of the ASR systems. A weight system may be also be used. The weight of each word may be used, given by module 207. The topic of the language model data is also taken into account. For example, it is very likely that the speech recognition systems may have different language models, and one ASR of the non-master user may have a better language model that is also similar to the topic of discussion. In this case, the word that was recognized on the non-master machine may have a higher weight than the decoded word from the master machine. The master machine may have a speaker dependent model while the other machines may have speaker independent models, all of which would directly affect the quality of the decoding. Examples of integration into non-master ASR systems from the master laptop will be given in later figures. Module 209 is the final decoder output. Module 210 prepares the summary of the entire decoded output of what was spoken, as per references XXXXXXX. Module 250 sends the final decoded data to a laptop(if needed) for transcription or editing.

Figure 3 describes the composition of laptop 100. The laptop contains a communication module 302 that allows the laptop to communicate with the server and other laptops. Module 301 represents a microphone connected to a driver 303 that is responsible for sending the speech data from the microphone into the speech recognition module 304, or into the communicator module for other laptops. The driver is also capable of receiving data from other laptops and sending it to the ASR 304. The ASR may also send part of decoded data 305 to the communication module, from what the ASR was able to decode, or other additional information (likelihood of the word, or information from other decoding modules). Examples of decoded data will be given in a later figure. The ASR 304 will be connected to different models like the speaker independent model 306, speaker dependent models 307, the master verification model 308 - checks that the master is speaking, and the ASR 304 will be able to do partial decoding and specific task recognition after receiving a partially decoded set of data from another ASR system on another computer(explained in a later figure).

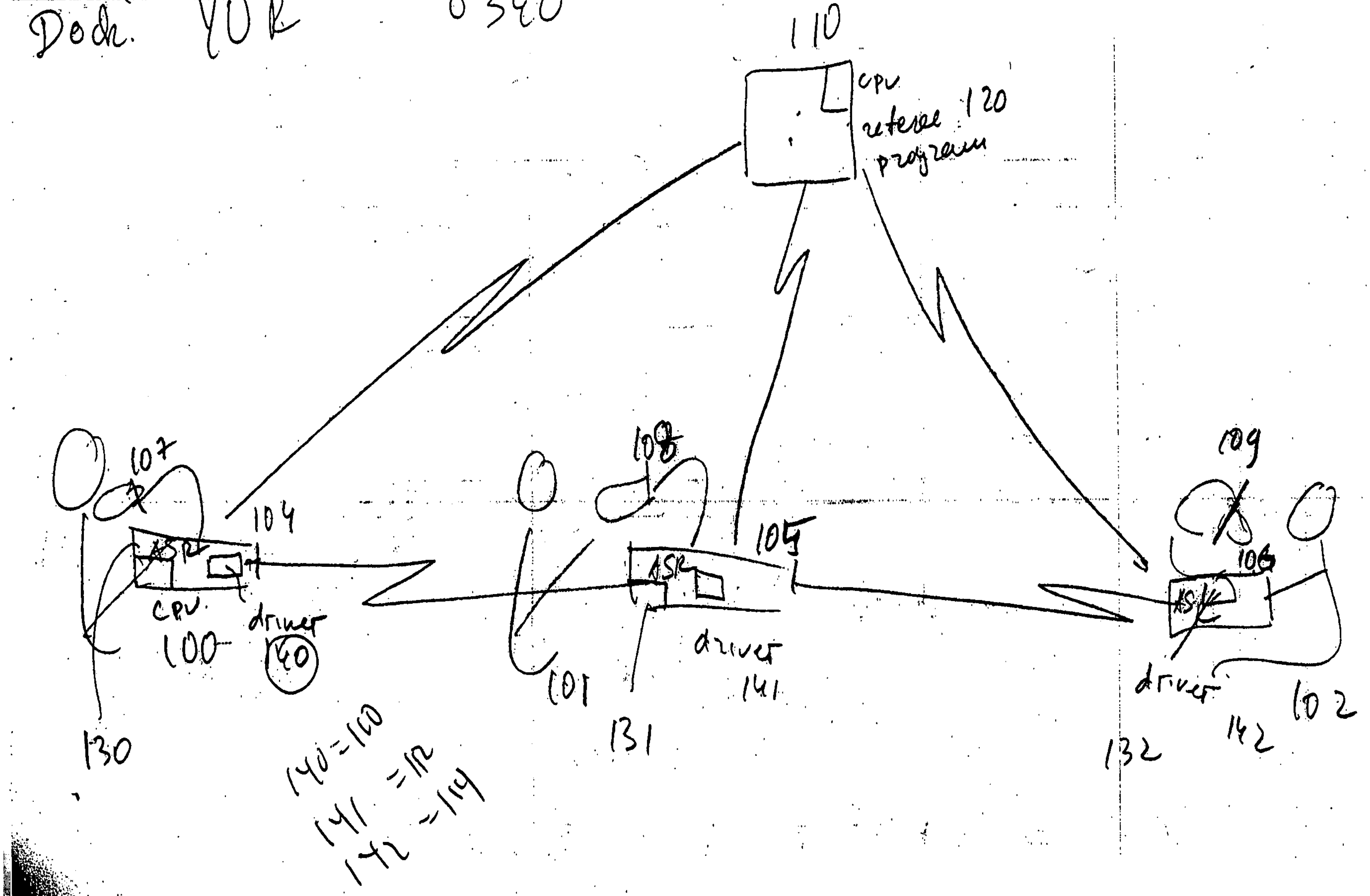
Figure 4 describes modules 305-decoder of data, and 309-the specific task recognizer. Module 400 represents an example of decoded data- text, words, and phonemes. Scores of words and phonemes are represented by module 401. Detailed matching of candidates may be done in module 402 as per reference XXXXXX; that produces detailed matching of candidates using specific models. When time-costly models are being decoded, module 402 produces a detailed list of candidates that may have a high chance of matching a particular set of acoustical data. Several words- W1, W2, and W3 may comprise and acoustic segment. Module 403 represents the fast matching of candidates, composed of words W1 and lists of words, that give an approximate method for finding candidates that are then narrowed by the fast match list. Acoustic data that was already processed by signal processing or by other feature vectors that may result from acoustic data(any process of speech recognition that results in a form of decoded data may send this data to the other speech recognitions).

Specific tasks done by the speech recognition are represented by module 410. For example, it may do detailed candidate decoding 411 using the words from modules 402 or 403. The candidates of words received by one speech recognition are sent over to another speech recognition where our own model is used to provide speech recognition. Similarly, phonetic sets 413 may be used by a speech recognitions system. The phonetic sets may change in each different ASR decoder. Depending on which phonetic set is used, the decoded result may be different. Different language models decoders, and different adaptation schemes (414 and 415) may be used. In other words, specific task recognition begins working from the module that represents the type of data that it received. If data was sent after fast matching, then it continues fast match in the present ASR system. If the data was sent after detailed match decoding, it uses the segment of data that was done after detailed match decoding.

Figure 5 gives and example of how the Integrator of all decoding data 208 functions. If we assume that the integrator received the 5 words from speech recognition, W1 with weight alpha1, W1 with weight alpha 2, W2 with weight alpha, W1 with weight alpha 4, and W2 with weight alpha 5(module 500). Module 501 compares if the weights of word W1 (alpha1+alpha2+alpha4) greater than or equal to the weights of word W2(alpha 3 + alpha 5). If the weight of W1 is greater than or equal to the weight of W2, then the system assumes that word W1 was said. If not, then the systems decides that word W2 was said. This scheme is one example of how the data may be integrated. Alpha 1,2,3,4, and 5 are the weights received from module 207 that gives the words a confidence score that may be based on topic reference.

Figure 6 is a block scheme of the invention methodology while also explaining the master verification system 308. Module 600 checks if the volume of the acoustic data greater than the threshold(module 308). If the volume is greater than a given threshold then 601 performs speaker verification for master. It checks if the master is speaking 602. If it is the master speaking, then speech recognition is performed in this machine. If the volume is less than a given threshold than the voice does not belong to the master and must have come from another source. If the volume is less than a given threshold then the system checks that the voice data belongs to a master in another computer via module 640-Is other voice master data available on another laptop? If yes, than it gets that data from the other master machine 603. It may also be that the data was background noise, and therefore not appear as a master voice on any of the other laptops. This is usually checked in module 601 that performs speaker verification-it may realize that the voice data does not belong to a speaker and was, in fact background noise. After the voice data is received from the master machine, the local machine assists in the decoding of the voice data from the master machine. Module 606 sends the data to a server where it is integrated. After this the data is sent to the summarator 607 that prepares a summary of the entire meeting, or it is sent to 608 the laptops that requested this information, for transcription or editing.

RBM
Dodd. YOR 0340



YOR 9 520

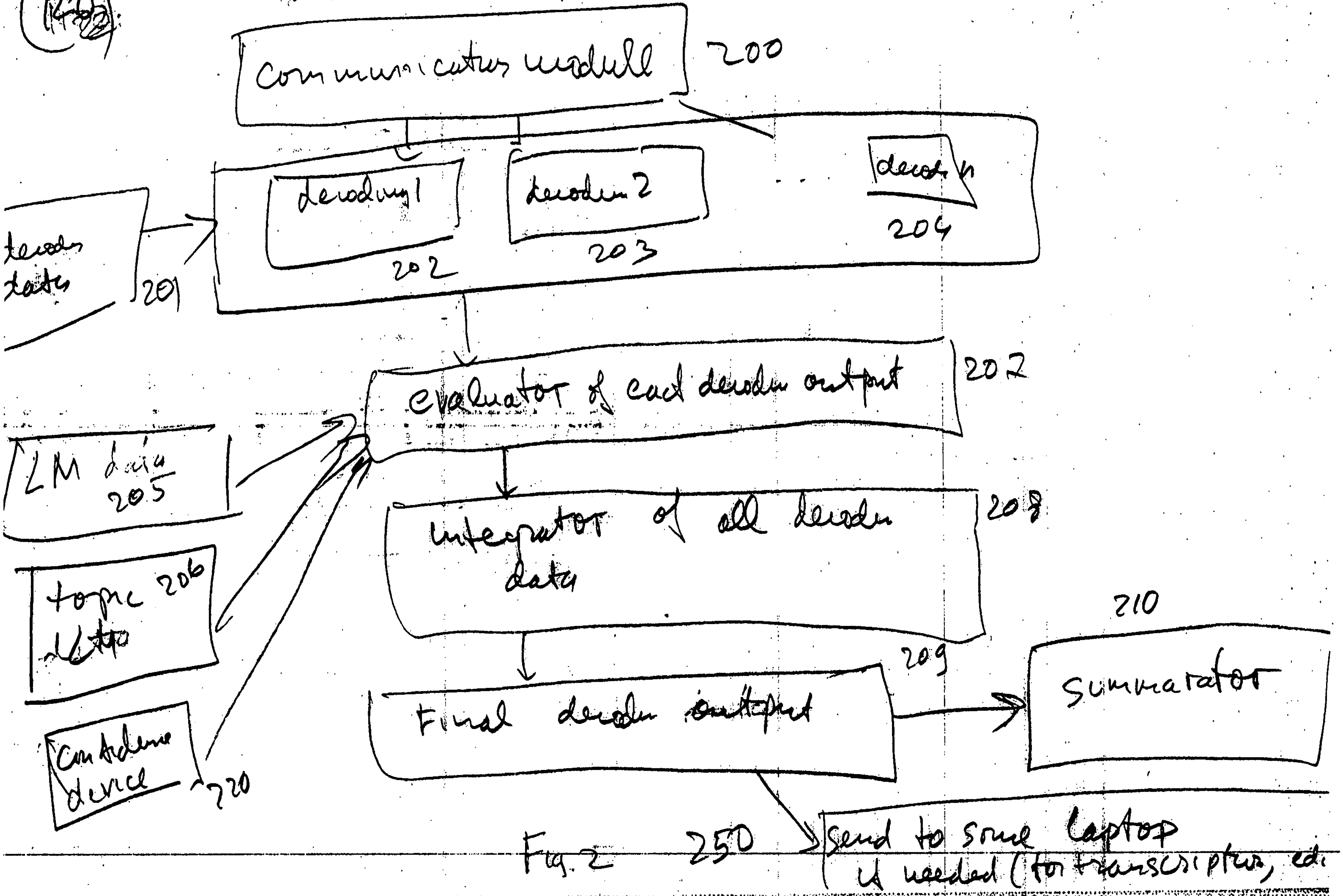


Fig. 2

(100) YOR9 0540

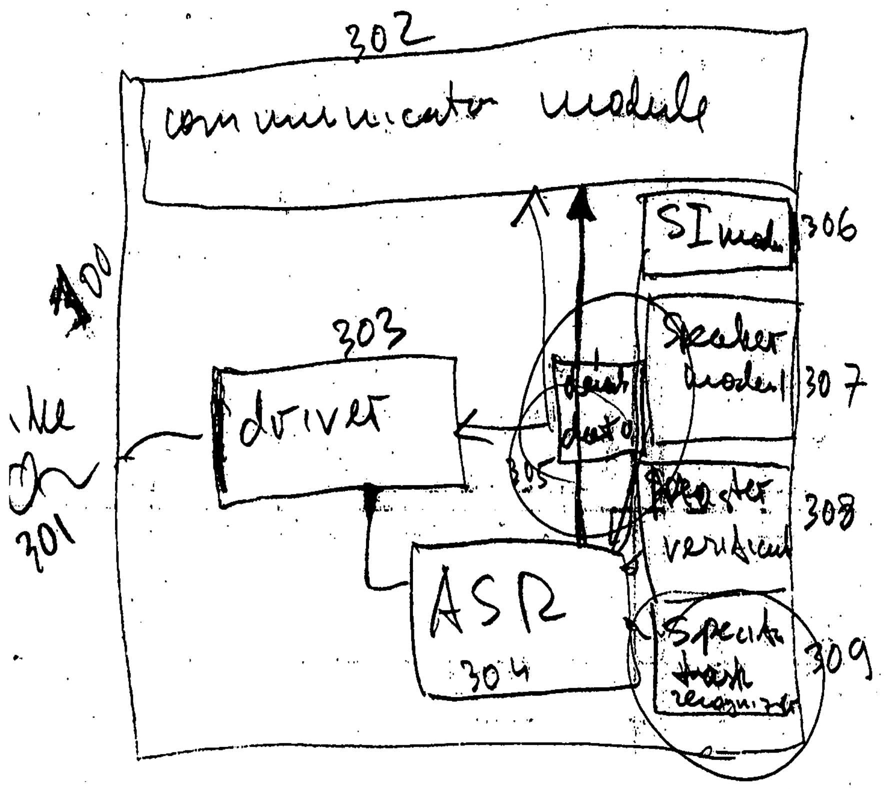


Fig. 3

(305
309)

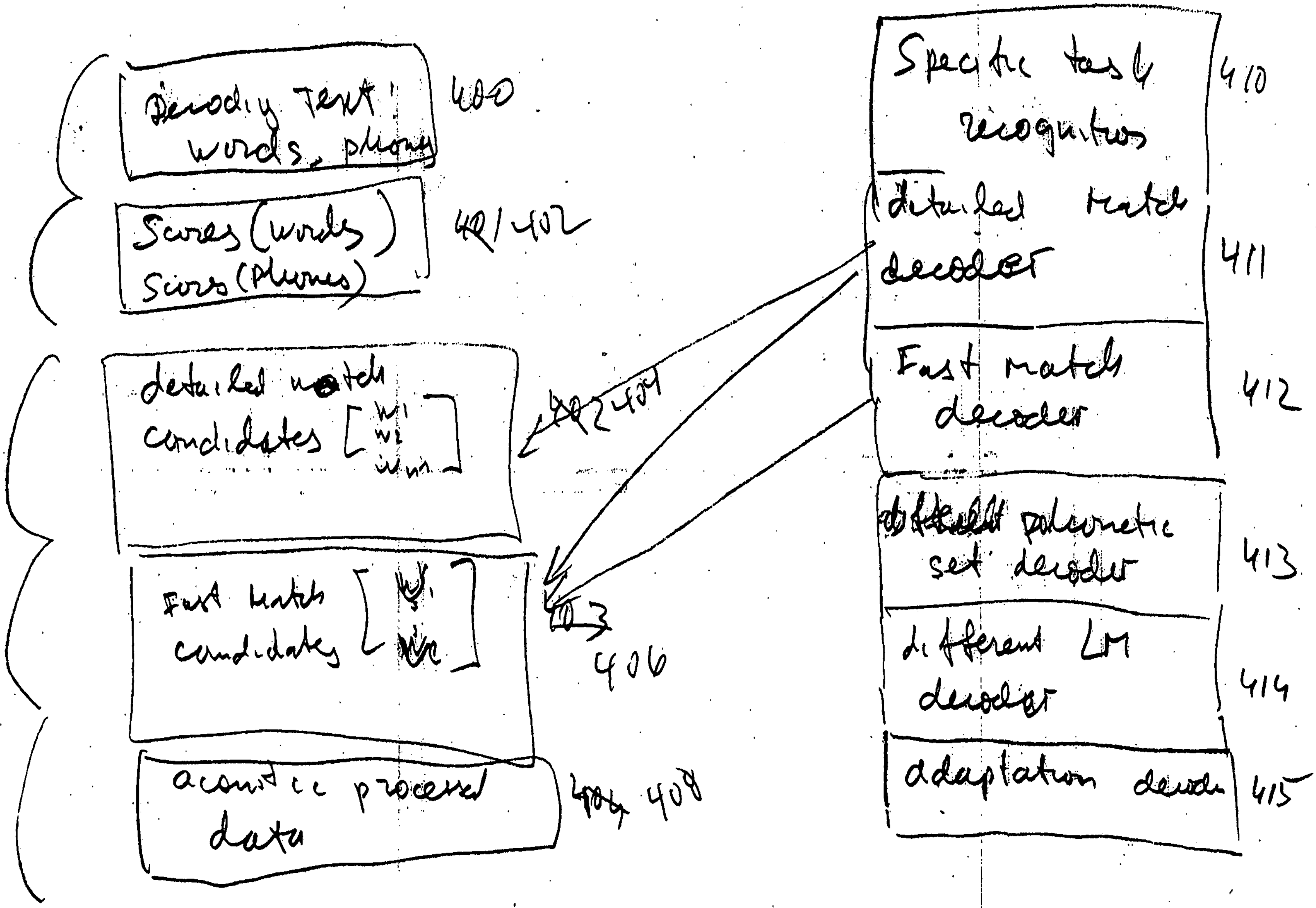

Decoding Text: words, phones
400
Scores (words)
Scores (phones)
401/402
detailed match candidates [w1 w2 wm]
~~403~~ 404
Fast match candidates [w1 wk]
~~403~~
406
acoustic processed data
~~404~~ 408
Specific task recognitors
410
detailed match decoder
411
Fast match decoder
412
~~different~~ paleonetic set decoder
413
different LM decoder
414
adaptation decoder
415

Fig.4